A Derivation of Back Propagation Through Time

Pengyu He
Johns Hopkins University
Baltimore, MD 21218
pyhe@jhu.edu

Abstract

We provide a derivation of Back Propagation Through Time for a neural language model.

1 Derivation

We use the notations below:

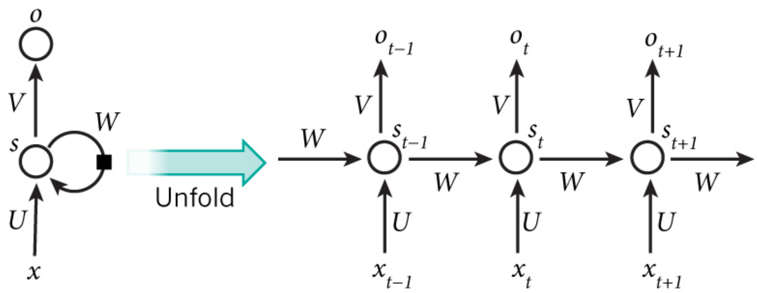

Figure 1: Data projection of top two CCA direction

$$\boldsymbol{x} \in R^V \quad \boldsymbol{o} \in R^V$$

$$\boldsymbol{s} \in R^m \text{(m is the dimension of hidden layer)}$$

$$\boldsymbol{U} \in R^{V\times m} \quad \boldsymbol{V} \in R^{m\times V} \quad \boldsymbol{W} \in R^{m\times m}$$

Suppose the nonlinear function is f at the output layer and F at hidden layer

We define the quadric loss function as

$$L = \frac{1}{2}\sum_{t=1}^{N}(\boldsymbol{d}^t - \boldsymbol{o}^t)^2 = \frac{1}{2}\sum_{t=1}^{N}\sum_{k=1}^{V}(d_k^t - o_k^t)^2 \tag{1}$$

In which $\boldsymbol{d}^t$ is the correct output, and $\boldsymbol{o}^t$ is the actual output. They are all distribution over the vocabularies so they have dimension of V

We first calculate the derivative of $\frac{\partial L}{\partial \boldsymbol{V}}$

$$\frac{\partial L}{\partial V_{ij}} = \sum_{t=1}^{N}\sum_{k=1}^{V}(o_k^t - d_k^t)\frac{\partial o_k^t}{\partial V_{ij}} \tag{2}$$

We suppose $\boldsymbol{g}^t = \boldsymbol{V}\boldsymbol{s}^t$, so $g_s^t = \sum_{r=1}^{m} V_{sr} s_r^t$

So

$$\frac{\partial o_k^t}{\partial V_{ij}} = \sum_{s=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}\frac{\partial g_s^t}{\partial V_{ij}} = \frac{\partial o_k^t}{\partial g_i^t}s_j^t = \frac{\partial f_k(g_1^t, g_2^t, \cdots, g_v^t)}{\partial g_i^t}s_j^t \tag{3}$$

Then we discuss $\frac{\partial L}{\partial W_{ij}}$

$$\frac{\partial L}{\partial W_{ij}} = \sum_{t=1}^{N}\sum_{k=1}^{V}(o_k^t - d_k^t)\frac{\partial o_k^t}{\partial W_{ij}} \tag{4}$$

$$\frac{\partial o_k^t}{\partial W_{ij}} = \sum_{s=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}\frac{\partial g_s^t}{\partial W_{ij}} = \sum_{s=1}^{m}\sum_{r=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{sr}\frac{\partial s_r^t}{\partial W_{ij}} \tag{5}$$

$\boldsymbol{s^t} = \boldsymbol{W}\boldsymbol{s}^{t-1} + \boldsymbol{U}\boldsymbol{x}^t$, $s_i^t = \sum_{p=1}^{m} W_{ip}s_p^{t-1} + \sum_{p=1}^{V} U_{ip}x_p^t$

So

$$\begin{aligned}\sum_{s=1}^{m}\sum_{r=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{sr}\frac{\partial s_r^t}{\partial W_{ij}} &= \sum_{s=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{si}\frac{\partial s_i^t}{\partial W_{ij}} \\ &= \sum_{s=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{si}\frac{\partial \sum_{p=1}^{m} W_{ip}s_p^{t-1}}{\partial W_{ij}} \\ &= \sum_{s=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{si}\left(s_j^{t-1} + \sum_{p=1}^{m} W_{ip}\frac{\partial s_p^{t-1}}{\partial W_{ij}}\right)\end{aligned} \tag{6}$$

Which can be recursively calculated.

Finally we discuss $\frac{\partial L}{\partial U_{ij}}$

$$\frac{\partial L}{\partial U_{ij}} = \sum_{t=1}^{N}\sum_{k=1}^{V}(o_k^t - d_k^t)\frac{\partial o_k^t}{\partial U_{ij}} \tag{7}$$

$$\begin{aligned}
\frac{\partial o_k^t}{\partial U_{ij}} &= \sum_{s=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}\frac{\partial g_s^t}{\partial U_{ij}} = \sum_{s=1}^{m}\sum_{r=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{sr}\frac{\partial s_r^t}{\partial U_{ij}} \\
&= \sum_{s=1}^{m}\sum_{r=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{sr}\frac{\partial\left(\sum_{p=1}^{m}W_{rp}s_p^{t-1} + \sum_{p=1}^{V}U_{rp}x_p^t\right)}{\partial U_{ij}} \\
&= \sum_{s=1}^{m}\sum_{r=1}^{m}\frac{\partial o_k^t}{\partial g_s^t}V_{sr}\left(\sum_{p=1}^{m}W_{rp}\frac{\partial s_p^{t-1}}{\partial U_{ij}} + \frac{\partial U_{rj}x_j^t}{\partial U_{ij}}\right)
\end{aligned} \tag{8}$$

And $\frac{\partial s_p^{t-1}}{\partial U_{ij}}$ can be recursively calculated.

References